Facilitating Research in Pathology using Natural Language Processing

Hua Xu, MS, Carol Friedman, PhD
Department of Medical Informatics, Columbia University

Abstract. Clinical research projects frequently rely on manual extraction of information from pathology reports, which is a costly and time-consuming process. This paper describes use of a natural language processing (NLP) system to automatically extract and structure information in textual pathology reports that is needed for clinical research.

Background. Many ongoing clinical research projects, such as projects involving studies associated with cancer [1], involve capturing information in pathology reports so that the information can be used to determine the eligibility of recruited patients for the study and to provide other information, such as cancer prognosis. Most of the data are usually extracted and entered manually because pathology reports consist primarily of unstructured free-text format and thus the information they contain are not useful for other automated processes that need to reliably access the information. NLP offers an alternative to automated coding, but is very challenging to develop. MedLEE [2] (Medical Language Extraction and Encoding System), a natural-language processor, has been successfully developed for automated encoding of the information content of text documents, including discharge summary, radiology, and pathology reports, but surgical pathology reports present a number of unique challenges, which this paper will describe. This paper will also describe the methods that were used to solve some of the problems. One of cancer research projects that we are collaborating with involves assessing ethnic disparities in breast cancer outcomes. For this project, variables such as the number of lymph nodes assessed, how many were positive, and the grade of the tumor, are needed. Additionally, information associated with DNA analysis and expression levels of estrogen and progesterone receptors and HER2-NEU are also important.

Methods. A sample of pathology reports associated with breast biopsies was retrieved from the data repository at Columbia-Presbyterian Medical Center (CPMC), and an analysis was performed. The analysis included several different steps. The first step determined the overall structure of the reports and identified the information types in each section. For example, there is usually a section called specimen, which often tell us the tissue type and/or body location of the target. The second step identified findings needed for the particular research project. The third step involved analyzing the sentences containing the findings and extending MedLEE's general schema for representing their structure. To adapt MedLEE so that it would recognize the new types of information, which were primarily genotypic concepts, new lexical entries were created. To minimize the modification to MedLEE, a pre-processing program was also developed to transform the reports into a format that MedLEE could process more accurately. The last step, which has not been developed yet, will consist of a post-processing program that will transform the data needed for the cancer registry so that the data can be entered directly into the registry database.

Challenges. A free-text pathology report is difficult to process directly by a natural language processor (NLP) for several reasons. One reason is that it contains tabular data and is missing punctuation marks in many places. For example, in Figure 1(a), instead of a period, a space is used to separate the value of the finding *DNA INDEX* from the next finding *S-PHASE*, and there is no period following *S-PHASE*. MedLEE depends on end of sentence markers to identify well-formed sentences. Therefore, the pre-processor adds periods at the appropriate places. Another difficulty is that some reports have explanations that are interspersed with findings and the information they contain should not be considered findings. Shown in Figure 1(b) is an explanation associated with the interpretation of *DNA INDEX* results. Our preprocessor identifies and marks the explanation statements so that they will be ignored. A third difficulty is that a pathology report often contains multiple specimens, and findings associated with each specimen are mentioned throughout the report. Figure 2 illustrates an example. It is difficult for an NLP system to associate findings with particular specimens, especially since the specimens are not uniformly identified in the reports. Our pre-processor links findings to the appropriate specimen by creating separate reports so that findings are associated with the appropriate specimen.

(**a**) DNA INDEX 1.03 S-PHASE 19.4%

(**b**) The DNA INDEX is the ratio of average… A DNA index of 1.0(0.9-1.1) indicates a diploid cell population. A DNA index of 2.0…

Figure 1

Specimen:
A: Breast, right,… B: Breast, left,…

Microscopic Description:
Slides A from the right breast …
Slides B from the left breast…

Figure 2

Conclusion. We have identified how information in pathology reports that is troublesome for an NLP system can be captured accurately by using a preprocessor to adjust the text and eliminate some of the difficulties. This project aims to demonstrate that an existing NLP system can be used to facilitate clinical research in clinical pathology reports with only minor modification.

References

1. Rosenberg DJ, Neugut AI, Ahsan H, Shea S. Diabetes mellitus and the risk of prostate cancer. Cancer Invest. 2002; 20(2): 157-65.
2. Friedman C, Alderson PO, Austin JH, Cimino JI, Johnson SB. A General Natural-language Text Processor for Clinical Radiology. J Am Med Informatics Assoc 1994; 1: 161-74.